

2A

(19)日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開平8-328760

(43)公開日 平成8年(1996)12月13日

(51)Int.Cl. ⁴	識別記号	庁内整理番号	F I	技術表示箇所
G 0 6 F 3/06	5 4 0		G 0 6 F 3/06	5 4 0
13/10	3 4 0	7368-5E	13/10	3 4 0 A

審査請求 未請求 請求項の数14 F D (全 21 頁)

(21)出願番号 特願平7-158370

(22)出願日 平成7年(1995)6月1日

(71)出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72)発明者 角田 仁

東京都国分寺市東恋ヶ窪一丁目280番地

株式会社日立製作所中央研究所内

(72)発明者 大山 光男

東京都国分寺市東恋ヶ窪一丁目280番地

株式会社日立製作所中央研究所内

(72)発明者 高本 良史

東京都国分寺市東恋ヶ窪一丁目280番地

株式会社日立製作所中央研究所内

(74)代理人 弁理士 矢島 保夫

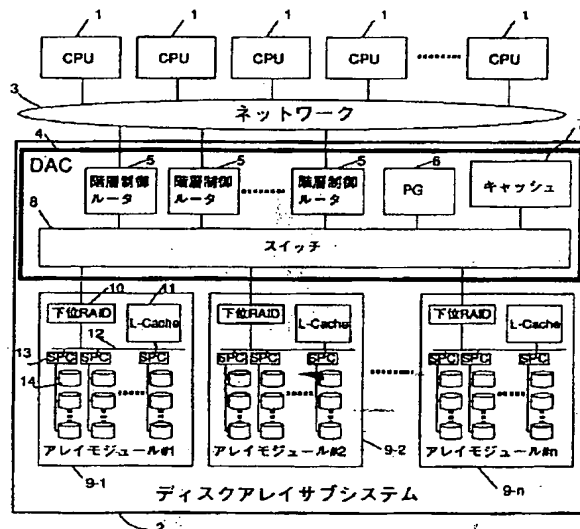
最終頁に続く

(54)【発明の名称】 ディスクアレイ装置

(57)【要約】

【目的】複数のディスクアレイモジュール間でディスクアレイ制御を実現する場合において、ディスクアレイ制御装置の基板上の実装の制約を排除し、制御を簡単にし、転送速度を向上させることのできるディスクアレイ装置を提供することを目的とする。

【構成】複数のディスクアレイモジュール間でディスクアレイ制御を実現する場合、従来のように複数のディスクアレイ制御装置をマザーボードに接続し、このマザーボード内の配線によるバスでこれらの複数のディスクアレイ制御装置間を制御するのではなく、並列動作が可能なクロスバ方式などのスイッチ手段により複数のディスクアレイモジュールを接続し、このスイッチのルーティング制御により複数のディスクアレイモジュール間によるディスクアレイ制御を実現する。



【特許請求の範囲】

【請求項1】上位装置に接続され、複数台のディスクアレイモジュール間でディスクアレイ制御を行うディスクアレイ装置であって、

上位装置から発行された読み出しまたは書き込み要求を受け付けるルータと、

各々が独立したディスクアレイ装置として内部でディスクアレイ制御を行っている複数台のディスクアレイモジュールと、

上記ルータ、および上記複数台のディスクアレイモジュールを各ポートに接続するとともに、それら各ポート間の接続を行うスイッチ手段とを備え、

上記ルータにより上記スイッチ手段の各ポート間の接続を制御することにより、上記複数台のディスクアレイモジュール間でディスクアレイ制御を行うことを特徴とするディスクアレイ装置。

【請求項2】上位装置に接続され、複数台のディスクアレイモジュール間でディスクアレイ制御を行うディスクアレイ装置であって、

上位装置から発行された読み出しまたは書き込み要求を受け付けるルータと、

パリティを生成するためのパリティ生成手段と、

各々が独立したディスクアレイ装置として内部でディスクアレイ制御を行っている複数台のディスクアレイモジュールと、

上記ルータ、上記パリティ生成手段、および上記複数台のディスクアレイモジュールを各ポートに接続するとともに、それら各ポート間の接続を行うスイッチ手段とを備え、

上記ルータにより上記スイッチ手段の各ポート間の接続を制御することにより、上記複数台のディスクアレイモジュール間でディスクアレイ制御を行うことを特徴とするディスクアレイ装置。

【請求項3】さらに、前記スイッチ手段のポートにキャッシュメモリを接続し、前記上位装置から発行された読み出しまたは書き込み要求に対して読み出しまたは書き込みを行うべきディスクアレイモジュールが接続されているポートを認識するためのルーティングテーブルを前記キャッシュメモリに記憶しておき、

前記ルータは、前記ルーティングテーブルを用いて、前記上位装置から発行された読み出しまたは書き込み要求に対して読み出しまたは書き込みを行うべきディスクアレイモジュールが接続されているポートを認識し、該ポートに接続されているディスクアレイモジュール間でディスクアレイ制御を行う請求項1または2に記載のディスクアレイ装置。

【請求項4】前記ルーティングテーブルを、前記キャッシュメモリ内に設ける代わりに、前記ルータから前記スイッチ手段を介さずに直接アクセスできるメモリ内に設けた請求項3に記載のディスクアレイ装置。

【請求項5】前記ルータは、前記上位装置からの書き込み要求を受け付けたとき、該書き込みデータを前記スイッチ手段を介して前記パリティ生成手段に転送し、前記パリティ生成手段において書き込みデータの分割およびパリティの生成を行い、該分割したデータおよび生成したパリティを、前記ルーティングテーブルを用いて認識された複数のディスクアレイモジュールに対してそれぞれ転送して書き込む請求項3または4に記載のディスクアレイ装置。

【請求項6】前記ルータは、前記上位装置からの読み出し要求を受け付けたとき、前記ルーティングテーブルを用いて、読み出すべき分割されたデータが格納されている複数のディスクアレイモジュールを認識し、前記スイッチ手段を介して該複数のディスクアレイモジュールと前記パリティ生成手段とを接続し、該複数のディスクアレイモジュールから前記パリティ生成手段へと分割されたデータを読み出し、前記パリティ生成手段は該読み出された分割されたデータを結合し、該結合したデータを前記パリティ生成手段から前記スイッチ手段および前記ルータを介して前記上位装置に転送する請求項3または4に記載のディスクアレイ装置。

【請求項7】前記ルータは、前記上位装置からの読み出し要求を受け付けた場合、読み出すべき分割されたデータが格納されている複数のディスクアレイモジュールのうちの何れかに障害が発生していたときは、前記ルーティングテーブルを用いて、読み出すべき分割されたデータ（障害が発生しているディスクアレイモジュールに格納されているデータを除く）が格納されている複数のディスクアレイモジュールおよびパリティが格納されているディスクアレイモジュールを認識し、前記スイッチ手段を介してそれらのディスクアレイモジュールと前記パリティ生成手段とを接続し、それらのディスクアレイモジュールから前記パリティ生成手段へと分割されたデータおよびパリティを読み出し、前記パリティ生成手段は該読み出された分割されたデータおよびパリティから障害が発生したディスクアレイモジュール内のデータを回復し、該回復したデータと読み出された分割されたデータを結合し、該結合したデータを前記パリティ生成手段から前記スイッチ手段および前記ルータを介して前記上位装置に転送する請求項6に記載のディスクアレイ装置。

【請求項8】前記ルータは、前記上位装置からの書き込み要求を受け付けたとき、前記ルーティングテーブルを用いて、該書き込みデータに対応する旧データおよび旧パリティが格納されているディスクアレイモジュールを認識し、該旧データおよび旧パリティを前記スイッチ手段を介して前記パリティ生成手段へと読み出し、前記パリティ生成手段は、読み出された旧データおよび旧パリティと書き込みデータとを用いて新パリティを生成し、該書き込みデータおよび生成した新パリティを、旧デー

たおよび旧バリティを読み出したディスクアレイモジュールに転送して書き込む請求項3または4に記載のディスクアレイ装置。

【請求項9】前記ルータは、前記上位装置からの読み出し要求を受け付けたとき、前記ルーティングテーブルを用いて、読み出すべきデータが格納されているディスクアレイモジュールを認識し、該ディスクアレイモジュールから読み出したデータを、前記スイッチ手段および前記ルータを介して前記上位装置に転送する請求項3または4に記載のディスクアレイ装置。

【請求項10】前記ルータは、前記上位装置からの読み出し要求を受け付けた場合、読み出したいデータが格納されているディスクアレイモジュールに障害が発生していたときは、前記ルーティングテーブルを用いて、当該読み出したいデータが作成に関与したバリティが格納されているディスクアレイモジュールおよび該バリティの作成に関与した当該読み出したいデータ以外のデータが格納されているディスクアレイモジュールを認識し、前記スイッチ手段を介してそれらのディスクアレイモジュールと前記バリティ生成手段とを接続し、それらのディスクアレイモジュールから前記バリティ生成手段へとバリティおよび該バリティの作成に関与したデータを読み出し、前記バリティ生成手段は該読み出されたバリティおよび該バリティの作成に関与したデータから障害が発生したディスクアレイモジュール内のデータを回復し、該回復したデータを前記バリティ生成手段から前記スイッチ手段および前記ルータを介して前記上位装置に転送する請求項9に記載のディスクアレイ装置。

【請求項11】前記上位装置が前記ルータに直接接続されており、前記ルータは、前記上位装置からの読み出しまたは書き込み要求を直接受け付ける請求項1または2に記載のディスクアレイ装置。

【請求項12】前記上位装置が前記スイッチ手段のポートに直接接続されており、前記ルータは、前記スイッチ手段を介して前記上位装置からの読み出しまたは書き込み要求を受け付ける請求項1または2に記載のディスクアレイ装置。

【請求項13】前記複数台のディスクアレイモジュールのうち任意の数のディスクアレイモジュールを、ディスクアレイ装置の本体とは別の筐体とした請求項1または2に記載のディスクアレイ装置。

【請求項14】請求項1または2のディスクアレイ装置のうち、複数台のディスクアレイモジュール以外の部分をディスクアレイ制御装置と呼ぶとき、複数のディスクアレイ制御装置で共通の複数台のディスクアレイモジュールを共有する請求項1または2に記載のディスクアレイ装置。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は、コンピュータシステム

における2次記憶装置に関し、特に高性能な入出力動作を可能とするディスクアレイ装置に関する。

【0002】

【従来の技術】現在のコンピュータシステムにおいては、CPU（中央処理装置）などの上位側が必要とするデータは2次記憶装置に格納され、CPUなどが必要とするときに応じて2次記憶装置に対してデータの書き込みおよび読み出しを行っている。この2次記憶装置としては、一般に不揮発な記憶媒体が使用され、代表的なものとして磁気ディスク装置（以下、ドライブとする）や、光ディスクなどがあげられる。

【0003】近年高度情報化に伴い、コンピュータシステムにおいて、この種の2次記憶装置の高性能化が要求されてきた。その一つの解として、多数の比較的容量の小さなドライブにより構成されるディスクアレイが考えられている。

【0004】公知の文献として、「D.Patterson,G.Gibson,and R.H.Kartz;A Case for Redundant Arrays of Inexpensive Disks(RAID),in ACM SIGMOD Conference,Chicago,IL,(June1988)」がある。この文献においては、全く同じデータを別のドライブに二重化して格納するミラーリングを行うディスクアレイ（レベル1）と、データを分割して並列に処理を行うディスクアレイ（レベル3）と、データを分散して独立に扱うディスクアレイ（レベル4、5）について、その性能および信頼性の検討結果が報告されている。レベル4は、レベル5において論理グループを構成するドライブに分散しているバリティを、1台のバリティのみを格納するドライブにまとめたものである。現在この論文に書かれている方式が、最も一般的なディスクアレイと考えられている。

【0005】ここで、レベル3、レベル4、およびレベル5のそれぞれについて、簡単に説明しておく。

【0006】まず、レベル3のディスクアレイについて簡単に説明する。ディスクアレイに格納するデータ#1として、例えば「001010101011・・・」を想定し、このデータ#1とバリティを格納するためのディスクとしてディスク#1～#5が設けられているとする。レベル3では、ディスク#1にデータ#1の第1ビット「0」を、ディスク#2にその次の第2ビット「0」を、ディスク#3に第3ビット「1」を、ディスク#4に第4ビット「0」を、順次格納し、格納された「0010」に対するバリティをディスク#5に格納する。そして、次に同様に、引き続きビット「1」、「0」、「1」、「0」を順次ディスク#1～#4に格納し、そのバリティを#5に格納してゆく。

【0007】レベル4では、データとバリティを格納するためのディスクとしてディスク#1～#5が設けられた場合、データ#1、#5、・・・がディスク#1に、データ#2、#6、・・・がディスク#2に、データ#3、#7、・・・がディスク#3に、データ#4、#

8、・・・がディスク#4に、それぞれ格納される。そして、例えば、データ#1が「01・・・」、データ#2が「00・・・」、データ#3が「11・・・」、データ#4が「00・・・」であるとする、各データの先頭ビットを並べた「0010」に対するパリティをパリティ専用として指定されたディスク#5の先頭ビットとして格納し、次に各データの2番目のビット「1010」に対するパリティをディスク#5の2番目のビットとして格納し、以下同様にしていく。そして、データ#5～#8のデータ組に対するパリティデータを、ディスク#5に2番目のパリティデータとして、格納するようにしてゆく。

【0008】レベル5は、レベル4のようにパリティ専用のディスクを決めず、データ#1をディスク#1に、データ#2をディスク#2に、データ#3をディスク#3に、データ#4をディスク#4にそれぞれ格納し、データ#1～#4のデータ組に対するパリティデータをディスク#5に格納し、次いで、データ#5をディスク#2に、データ#6をディスク#3に、データ#7をディスク#4に、データ#8をディスク#5にそれぞれ格納し、データ#5～#8のデータ組に対するパリティデータをディスク#1に格納し、次いで、データ#9をディスク#1に、データ#10をディスク#3に、データ#11をディスク#4に、データ#12をディスク#5にそれぞれ格納し、データ#9～#12のデータ組に対するパリティデータをディスク#2に格納する、とようにしてゆく。要するに、パリティデータを分散して格納するものである。

【0009】上記文献に記載されたタイプのディスクアレイでは、大型大容量のドライブを、多数の比較的容量の小さなドライブで構成し、データを分散して格納するため、読み出し/書き込み要求が増加してもディスクアレイの複数のドライブで分散して処理することが可能となり、読み出し/書き込み要求が待たされることが減少する。

【0010】次に、ディスクアレイにおけるパリティについて説明する。ディスクアレイは従来の大容量のドライブを、比較的容量の小さな多数のドライブで構成するため、部品点数が増加し障害が発生する確率が高くなる。このため、ディスクアレイでは、上述したようにパリティを用意している。パリティを用意することにより、データを格納したドライブに障害が発生した場合でもその障害ドライブ内のデータを復元することが可能となる。ディスクアレイでは、データからパリティを作成しデータと同様にドライブに格納しておく。このとき、パリティは、パリティの作成に関与したデータとは別のドライブに格納される。

【0011】特開平6-119120号公報は、ディスクアレイにおけるデータの更新時に、キャッシュメモリ内に当該データが存在する場合は、キャッシュメモリ内

の当該データを更新し、所定回数の更新が行われた場合、ディスクアレイ内の冗長データ(パリティ)を更新する方法について開示している。この特開平6-119120号公報では、ディスクアレイ制御装置の構成例が図2および図3に示されている。ここで示されている構成例は、現在最も一般的な構成方法と考えられる。具体的には、ディスクアレイ制御装置には制御手段として動作するMPU(マイクロプロセッサ)が設けられ、MPUからの内部バスに対し、処理プログラムを格納したROM(リード・オンリ・メモリ)、制御記憶等として用いられるRAM(ランダム・アクセス・メモリ)、キャッシュ制御部を介して接続したキャッシュメモリ、およびデータ転送バッファが設けられる。このように、ディスクアレイ制御装置内において、MPU、ROM、RAM、およびキャッシュ制御部は、これらが共有で使用する内部バスにより相互に接続されている。

【0012】また、別の公知例としては特開平6-242887号公報がある。この公知例では、CPUとディスクアレイ(ディスクアレイコントローラ+ドライブ)とをネットワークスイッチにより接続することで、ディスクアレイのI/O帯域幅とCPUのI/O性能を最適に整合させる方法について開示している。つまり、高性能かつ自由度の高いネットワークスイッチを介してCPUとディスクアレイとを接続することで、接続の自由度を向上させ、ネックを解消するようにしている。

【0013】

【発明が解決しようとする課題】現在のディスクアレイ制御装置は、特開平6-119120号公報に示されているように、制御手段として動作するMPUが設けられ、処理プログラムを格納したROM、制御記憶等として用いられるRAM、キャッシュ制御部を介して接続したキャッシュメモリ、およびデータ転送バッファは、内部バスを介してMPUに接続されている。このように、ディスクアレイ制御装置内では、MPUにおけるすべての制御が内部バスを介して行われる。この内部バスは、実質的にはMPU、ROM、RAM、およびキャッシュ等が実装されている基板上の配線である。この、基板上の配線による内部バスの性能は、内部バスの転送速度(MB/s)と内部バスのバス幅(B)との積で決定される。しかし、基板が大きくなり配線長が長くなると、転送速度を向上させることが困難になる。特に、今後複数のディスクアレイ制御装置間でディスクアレイ制御を実現しようとした場合、複数のディスクアレイ制御装置が接続されるマザーボードはかなり大きくなるため、転送速度を向上させることはさらに困難になる。

【0014】また、内部バスのバス幅を広げると基板内での実装やコネクタが大きくなるため、実装上の問題が生じる。

【0015】以上より、今後ディスクアレイ制御装置の性能を向上させるためには、このような基板内での配線

を使用する内部バスが性能のネックになるという問題を解決しなければならない。

【0016】本発明は、ディスクアレイ制御装置の基板上の実装の制約を排除し、制御を簡単にして転送速度を向上させることのできるディスクアレイ装置を提供することを目的とする。

【0017】

【課題を解決するための手段】上記目的を達成するため、請求項1に記載の発明は、上位装置に接続され、複数台のディスクアレイモジュール間でディスクアレイ制御を行うディスクアレイ装置であって、上位装置から発行された読み出しまたは書き込み要求を受け付けるルータと、各々が独立したディスクアレイ装置として内部でディスクアレイ制御を行っている複数台のディスクアレイモジュールと、上記ルータ、および上記複数台のディスクアレイモジュールを各ポートに接続するとともに、それら各ポート間の接続を行うスイッチ手段とを備え、上記ルータにより上記スイッチ手段の各ポート間の接続を制御することにより、上記複数台のディスクアレイモジュール間でディスクアレイ制御を行うことを特徴とする。

【0018】また、請求項2に記載の発明は、請求項1に記載の発明の構成に加えてパリティ生成手段をもスイッチ手段に接続するようにして、RAIDレベル4、5にも適用できる点を明らかにしたものである。

【0019】スイッチ手段の各ポート間の接続を制御することにより複数台のディスクアレイモジュール間でディスクアレイ制御を行うため、上位装置から発行された読み出しまたは書き込み要求に対して読み出しまたは書き込みを行うべきディスクアレイモジュールが接続されているポートを認識するためのルーティングテーブルを用いるとよい。ルーティングテーブルは、キャッシュメモリ内に設けてもよいし、スイッチ手段を介さずに直接アクセスできるメモリ内に設けてもよい。

【0020】請求項5に記載の発明は、請求項3または4に記載のディスクアレイ装置においてRAIDレベル3のディスクアレイ制御を行う際の書き込み要求に対する処理を明らかにしたものであり、前記上位装置からの書き込み要求を受け付けたとき、該書き込みデータを前記スイッチ手段を介して前記パリティ生成手段に転送し、前記パリティ生成手段において書き込みデータの分割およびパリティの生成を行い、該分割したデータおよび生成したパリティを、前記ルーティングテーブルを用いて認識された複数のディスクアレイモジュールに対してそれぞれ転送して書き込むことを特徴とする。

【0021】請求項6に記載の発明は、請求項3または4に記載のディスクアレイ装置においてRAIDレベル3のディスクアレイ制御を行う際の読み出し要求に対する処理を明らかにしたものであり、前記上位装置からの読み出し要求を受け付けたとき、前記ルーティングテ

ブルを用いて、読み出すべき分割されたデータが格納されている複数のディスクアレイモジュールを認識し、前記スイッチ手段を介して該複数のディスクアレイモジュールと前記パリティ生成手段とを接続し、該複数のディスクアレイモジュールから前記パリティ生成手段へと分割されたデータを読み出し、前記パリティ生成手段は該読み出された分割されたデータを結合し、該結合したデータを前記パリティ生成手段から前記スイッチ手段および前記ルータを介して前記上位装置に転送することを特徴とする。

【0022】請求項7に記載の発明は、さらにRAIDレベル3のディスクアレイ制御を行う際の読み出し要求で、読み出すべき分割されたデータが格納されている複数のディスクアレイモジュールのうちの何れかに障害が発生していたときの処理を明らかにしたものである。すなわち、前記ルータは、前記上位装置からの読み出し要求を受け付けた場合、読み出すべき分割されたデータが格納されている複数のディスクアレイモジュールのうちの何れかに障害が発生していたときは、前記ルーティングテーブルを用いて、読み出すべき分割されたデータ（障害が発生しているディスクアレイモジュールに格納されているデータを除く）が格納されている複数のディスクアレイモジュールおよびパリティが格納されているディスクアレイモジュールを認識し、前記スイッチ手段を介してそれらのディスクアレイモジュールと前記パリティ生成手段とを接続し、それらのディスクアレイモジュールから前記パリティ生成手段へと分割されたデータおよびパリティを読み出し、前記パリティ生成手段は該読み出された分割されたデータおよびパリティから障害が発生したディスクアレイモジュール内のデータを回復し、該回復したデータと読み出された分割されたデータを結合し、該結合したデータを前記パリティ生成手段から前記スイッチ手段および前記ルータを介して前記上位装置に転送することを特徴とする。

【0023】請求項8に記載の発明は、請求項3または4に記載のディスクアレイ装置においてRAIDレベル4、5のディスクアレイ制御を行う際の書き込み要求に対する処理を明らかにしたものであり、前記ルータは、前記上位装置からの書き込み要求を受け付けたとき、前記ルーティングテーブルを用いて、該書き込みデータに対応する旧データおよび旧パリティが格納されているディスクアレイモジュールを認識し、該旧データおよび旧パリティを前記スイッチ手段を介して前記パリティ生成手段へと読み出し、前記パリティ生成手段は、読み出された旧データおよび旧パリティと書き込みデータとを用いて新パリティを生成し、該書き込みデータおよび生成した新パリティを、旧データおよび旧パリティを読み出したディスクアレイモジュールに転送して書き込むことを特徴とする。

【0024】請求項9に記載の発明は、請求項3または

4に記載のディスクアレイ装置においてRAIDレベル4、5のディスクアレイ制御を行う際の読み出し要求に対する処理を明らかにしたものであり、前記ルータは、前記上位装置からの読み出し要求を受け付けたとき、前記ルーティングテーブルを用いて、読み出すべきデータが格納されているディスクアレイモジュールを認識し、該ディスクアレイモジュールから読み出したデータを、前記スイッチ手段および前記ルータを介して前記上位装置に転送することを特徴とする。

【0025】請求項10に記載の発明は、さらにRAIDレベル4、5のディスクアレイ制御を行う際の読み出し要求で、読み出すべきデータが格納されているディスクアレイモジュールに障害が発生していたときの処理を明らかにしたものである。すなわち、前記ルータは、前記上位装置からの読み出し要求を受け付けた場合、読み出したいデータが格納されているディスクアレイモジュールに障害が発生していたときは、前記ルーティングテーブルを用いて、当該読み出したいデータが作成に関与したパリティが格納されているディスクアレイモジュールおよび該パリティの作成に関与した当該読み出したいデータ以外のデータが格納されているディスクアレイモジュールを認識し、前記スイッチ手段を介してそれらのディスクアレイモジュールと前記パリティ生成手段とを接続し、それらのディスクアレイモジュールから前記パリティ生成手段へとパリティおよび該パリティの作成に関与したデータを読み出し、前記パリティ生成手段は該読み出されたパリティおよび該パリティの作成に関与したデータから障害が発生したディスクアレイモジュール内のデータを回復し、該回復したデータを前記パリティ生成手段から前記スイッチ手段および前記ルータを介して前記上位装置に転送することを特徴とする。

【0026】本発明では、上位装置からの読み出した書き込み要求はルータにより受け付けられ、ルータがスイッチ手段を制御してRAID制御を行う。この場合、上位装置とルータとを直接接続してもよいし、スイッチ手段を介して接続してもよい。

【0027】複数台のディスクアレイモジュールのうち任意の数のディスクアレイモジュールを、ディスクアレイ装置の本体とは別の筐体としてもよい。また、複数台のディスクアレイモジュール以外の部分をディスクアレイ制御装置と呼ぶとき、複数のディスクアレイ制御装置で共通の複数台のディスクアレイモジュールを共有するようにしてもよい。

【0028】

【作用】本発明によれば、複数のディスクアレイ制御装置間でディスクアレイ制御を実現する場合、従来のように複数のディスクアレイ制御装置をマザーボードに接続し、このマザーボード内の配線によるバスでこれらの複数のディスクアレイ制御装置間を制御するのではなく、例えば、並列動作が可能なクロスバ方式等のスイッチな

どのスイッチ手段により複数のディスクアレイモジュールを接続し、このスイッチ手段のルーティング制御により複数のディスクアレイモジュール間によるディスクアレイ制御を実現する。

【0029】複数のディスクアレイモジュール間によるディスクアレイ制御をクロスバ方式などのスイッチ手段のルーティング制御により実現することにより、実装を気にすることなく転送速度が向上する。すなわち、1本当りの転送速度は従来のバス方式と大差無いが、スイッチを中心にした場合、複数本のバスが並列に動作することが可能となる。また、スイッチを切り替えることでルーティングを行いディスクアレイ制御を行うことが可能となる。

【0030】

【実施例】以下、図面を用いて本発明の実施例を説明する。

【0031】（実施例1）

【0032】図1は、本発明に係る第1の実施例のハードウェア構成を示す。1はCPU、2はディスクアレイサブシステム、3はCPU1とディスクアレイサブシステムを結ぶネットワークである。

【0033】ディスクアレイサブシステム2は、図1に示すように、ディスクアレイ制御装置（DAC）4とn台のアレイモジュール9（9-1、9-2、…、9-n）で構成されている。DAC4は、複数の階層制御ルータ5とパリティ生成回路（PG）6とキャッシュメモリ7とスイッチ8で構成されている。階層制御ルータ5は、マイクロプロセッサで構成されている。階層制御ルータ5とパリティ生成回路（PG）6とキャッシュメモリ7とアレイモジュール9は、スイッチ8の各ポートに接続されている。

【0034】図2は、DAC4内のスイッチ8の内部構成を示している。n個（nは任意）の上位ポート20には、階層制御ルータ5やパリティ生成回路（PG）6やキャッシュメモリ7が接続される。m個（mは任意）の下位ポート21には、アレイモジュール9が接続される。

【0035】このスイッチ8では、上位、下位のポートにかかわらず、任意のポートの接続が可能である。

【0036】図3は、アレイモジュール9の内部構成を示す。アレイモジュール9は、下位RAID制御部（以下、単に下位RAIDと呼ぶ）10とローカルキャッシュメモリ11とドライブとのSCSI（small computer system interface）インターフェース制御を行うSPC12とが内部バス12に接続されて構成されている。各SPC12には、SCSIバスにより複数のドライブ14が接続されている。これらのドライブ14に対しては、SPC12のSCSIインターフェース制御により、データの読み出しまたは書き込みが行われる。

【0037】下位RAID10は、アレイモジュール9

の内部でのディスクアレイ制御を行うための下位RAID制御用マイクロプロセッサ(MP)18と、このMP18が使用するMP用メモリ15と、MP18上で動作するマイクロプログラムが格納されるブートROM16と、スイッチ8とのインターフェース制御を行うインターフェース制御回路17とで構成され、これらはすべて内部バス12に接続されている。インターフェース制御回路17は、線19により、スイッチ8の下位ポート21に接続される。

【0038】このアレイモジュール9は、これ単独でディスクアレイとして機能する。すなわち、スイッチ8の下位ポート21を経由して転送されてきたデータは、アレイモジュール9内のMP18により予め指定されたRAIDのレベルに従って当該ドライブ14に格納される。このアレイモジュール9内のRAIDの制御は、従来より一般に行われている制御方法と同じで構わない。

【0039】図1のディスクアレイサブシステム2では、下位側として従来のディスクアレイであるアレイモジュール9内でディスクアレイのRAID制御を行い、さらに上位側としてDAC4によりアレイモジュール9間でディスクアレイのRAID制御を行うことで、2段階のRAID制御を行うようにしている。本実施例の特徴は、このような2段階のRAID制御を行う際に、階層制御ルータ5によりスイッチ8の切り換えを制御することによって、上位側のDAC4におけるRAID制御を実現したことにある。以下に、その方法について説明する。

【0040】図4は、DAC4内の階層制御ルータ5が使用するルーティングテーブル22の一例を示す。ルーティングテーブル22は、階層制御ルータ5ごとに設けられ、DAC4内のキャッシュメモリ7に格納されている。DAC4内の各階層制御ルータ5は、CPU1から読み出しまたは書き込み要求が発行された場合、スイッチ8を介してキャッシュメモリ7内のルーティングテーブル22へアクセスする。

【0041】なお、キャッシュメモリ7のようなスイッチ8のポートに接続したメモリにルーティングテーブル22を格納する代わりに、各階層制御ルータ5から直接アクセス可能なメモリを用意し、そこにルーティングテーブル22を格納する方法もある。このようにした場合、各階層制御ルータ5から直接アクセス可能なルーティングテーブル22用のメモリを用意しなければならないが、CPU1から読み出しまたは書き込み要求が発行される毎にスイッチ8を介してキャッシュメモリ7内のルーティングテーブル22へアクセスする必要はなくなるので、スイッチ8を使用しなくてもよくなり、スイッチ8が性能のネックになる場合は有効である。

【0042】本実施例では、CPU1からの読み出しまたは書き込み要求はシーケンスとして送られてくる。シーケンスには、当該シーケンスを識別するためにIDが

付けられる。そして、ルーティングテーブル22を参照してこのシーケンスID23に対応するエントリを認識することにより、DAC4内の階層制御ルータ5はRAID制御を行う。シーケンスIDは、図4に示すように、当該処理の番号(シーケンス番号)と、読み出しまたは書き込みされるデータのアドレスで構成される。上位装置であるCPU1からは、このシーケンスIDに読み出しまたは書き込み(R/W)コマンドが付けられて送られてくる。

【0043】図4は、RAIDのレベル1のミラーリング(2重化)のRAID制御を行う場合に使用されるルーティングテーブル22を示す。従来技術の欄で説明した特開平6-242887号の技術では、CPUとディスクアレイとを接続するネットワークスイッチではRAID制御を行わないため、このネットワークスイッチにおけるルーティング制御においては、CPUから送出される一つのシーケンスに対し送信先のポートは1個であった。これに対し、本実施例では、DAC4において階層制御ルータ5がスイッチ8を切り替えることでRAID制御を実現する。このため、一つのシーケンスID23に対し、階層制御ルータ5がルーティングテーブル22を用いて複数の送信先ポート(例えば、図4の例では2重化を行うため2つの送信先ポート124と送信先ポート225)に変換する。

【0044】本実施例では、ディスクアレイサブシステム2の初期設定の段階で、DAC4およびアレイモジュールで行われるRAIDのレベルが設定される。つまり、DAC4においては、どのアレイモジュール9のグループにより、どのようなレベルのRAID制御が行われるのかを初期設定の段階で設定し、アレイモジュール9においても、どのドライブ14のグループにより、どのようなレベルのRAID制御が行われるのかを初期設定の段階で設定しておく。このため、ディスクアレイサブシステム2の使用開始時において、ルーティングテーブル22は、送信先ポート124および送信先ポート225のみ設定されており、シーケンスID23には何も登録されていない。そこで、このディスクアレイサブシステム2の使用を開始し、データの新規書き込みを行う場合は、ルーティングテーブル22のシーケンスID23に書き込むデータのシーケンスIDを登録し、ルーティングテーブル22上でこの登録を行ったシーケンスID23に対応する送信先ポート124および送信先ポート225に接続されているアレイモジュール9にデータを転送する。また、データを削除する場合は、ルーティングテーブル22上において削除するデータのシーケンスID23を削除し、同様に、アレイモジュール9内のテーブル上の当該シーケンスIDを削除する。

【0045】なお、通常に使用している際に、新たに新しいデータを書き込みたい場合は、上述したような新規にデータを書き込む場合と同様に、ルーティングテーブ

ル 22 上においてシーケンス ID 23 が登録されていない所を探し、シーケンス ID 23 が登録されていなければ、そこに書き込むデータのシーケンス ID を登録し、対応する送信先ポート 1 24 および送信先ポート 2 25 に接続されているアレイモジュール 9 にデータを転送する。このようにして DAC 4 からデータを受け取ったアレイモジュール 9 では、同様に初期設定の段階で設定しておいた RAID のレベルに従い、ドライブ 14 にデータを書き込む。

【0046】図 5 に、DAC 4 において階層制御ルータ 5 が RAID のレベル 1 (RAID1) の制御を行う際のフローチャートを示す。階層制御ルータ 5 は、マイクロプロセッサで構成されている。

【0047】まず、CPU 1 より読み出しまたは書き込み要求が発生し、シーケンス ID が付けられて、ディスクアレイサブシステム 2 の DAC 4 に発行されたとする (ステップ 26)。本実施例では、予め DAC 4 で行う RAID 制御のレベル (本例では RAID レベル 1) をユーザが指定しておいてある。そこで、CPU 1 から発行された、読み出しまたは書き込み要求を受け付けた階層制御ルータ 5 は、予め指定されている RAID のレベルを認識する (ステップ 27)。図 5 の例では、RAID1 の制御を行うように認識する。

【0048】次に、読み出しまたは書き込み要求を受け付けた階層制御ルータ 5 は、キャッシュメモリ 7 に格納されているルーティングテーブル 22 を検索する (ステップ 28)。図 5 の例では、RAID1 の制御を行うためルーティングテーブル 22 は図 4 のようになっている。そして、RAID1 のミラーリングの制御を DAC 4 の階層制御ルータ 5 が行うため、CPU 1 から発行された 1 個の読み出しまたは書き込み要求は、図 4 のルーティングテーブル 22 により、送信先ポート 1 24 と送信先ポート 2 25 の 2 個のポートに変換される (ステップ 29)。「変換」とは、具体的には、新規データ書き込みの場合は、図 4 のルーティングテーブル 22 からシーケンス ID 23 の欄が空きのエントリを探し、そこに書き込み要求のシーケンス ID を登録し、対応する送信先ポート 1 24 と送信先ポート 2 25 を認識することである。また、データの読み出しあるいは更新の場合は、読み出しまたは書き込み (更新) 要求のシーケンス ID を図 4 のルーティングテーブル 22 から探し、対応する送信先ポート 1 24 と送信先ポート 2 25 を認識することである。

【0049】階層制御ルータ 5 が複数 (2 個) の送信先の下位ポート 21 を認識したら、当該階層制御ルータ 5 が接続されている上位ポート 20 と当該下位ポート 21 を接続する (ステップ 30)。

【0050】このとき、当該アレイモジュール 9 が他の読み出しまたは書き込み要求で使用中の場合は、使用可能になるまで待つ (ステップ 31)。当該アレイモジュ

ール 9 が使用可能になったら、これらの当該アレイモジュール 9 が接続されている下位ポートに対し、読み出しまたは書き込み要求を発行する (ステップ 32)。つまり、階層制御ルータ 5 は、自分が接続されているスイッチ 8 の上位ポート 20 と当該下位ポート 21 (本例では 2 個の下位ポート) とを接続し、線 19 を介して当該アレイモジュール 9 に対し読み出しまたは書き込み要求を発行することになる。

【0051】このようにして読み出しまたは書き込み要求を受け取った当該アレイモジュール 9 では、アレイモジュール 9 内の MP 18 が、この要求の受付が可能かどうかを判断し、先に述べたようにアレイモジュール 9 内で独自に読み出しまたは書き込み処理を行い、CPU 1 とアレイモジュール 9 間でデータの読み出しまたは書き込みが可能であれば、線 19、当該下位ポート 21、当該上位ポート 20 を介し階層制御ルータ 5 に対し転送許可を発行する (ステップ 33)。この当該アレイモジュール 9 からの転送許可を受け取った階層制御ルータ 5 は、CPU 1 と当該アレイモジュール 9 との間でデータ転送の制御を行う (ステップ 34)。

【0052】上述の手順はデータの 2 重化を行う RAID1 の制御を行うものであるため、同じデータが別のアレイモジュール 9 内に 2 重化されて格納される。このため、読み出し時は、当該データが格納されている 2 個のアレイモジュール 9 のうちで早く処理可能な方から当該データを読み出す。したがって、詳しく言えば、上記ステップ 30 では 2 個の下位ポートのうちのどちらか一方と上位ポートを接続すればよい。一方、書き込み時は、両方のアレイモジュール 9 内に同じデータを転送し書き込む。

【0053】(実施例 2)

【0054】実施例 1 では 2 重化を行う RAID1 の制御を例に説明した。実施例 2 では、本発明を RAID のレベル 3 (RAID3) に適用した場合を説明する。従来技術の欄で説明したが、RAID3 では一つのデータを複数に分割しこれらを複数の別のドライブに並列に書き込む。このとき、ドライブの障害に対する信頼性を向上させるため、分割したデータからパリティを作成し、データとは別のドライブに格納する。読み出す場合は、逆に複数のドライブから分割されたデータを並列に読み出し、結合して上位の CPU 1 へ転送する。

【0055】本実施例では、DAC 4 の階層制御ルータ 5 が複数のアレイモジュール 9 に対し、DAC 4 内のスイッチ 8 の切り替えを制御することにより、これらの複数のアレイモジュール 9 に対する RAID3 の制御を行う。そこで、以下にその方法を説明する。なお、全体のハードウェア構成、スイッチの内部構成、およびアレイモジュールの内部構成については、上述の実施例 1 の図 1、図 2、および図 3 と同じであるので説明は省略する。

【0056】図6は、RAID3のときのDAC4内の階層制御ルータ5が使用するルーティングテーブル22を示している。各階層制御ルータ5のルーティングテーブル22の使用方法是、上述した実施例1のRAID1の場合と同じである。

【0057】RAID3においても、ルーティングテーブル22におけるシーケンスID23に対応するエントリを認識することにより、DAC4内の階層制御ルータ5はRAID制御を行う。本実施例では、DAC4において階層制御ルータ5がスイッチ8を切り替えることでRAID制御を実現する。このため、一つのシーケンスID23に対し、階層制御ルータ5がルーティングテーブル22を用いて複数の送信先ポート（図6の例ではRAID3制御を行うためデータ用の4つの送信先ポート1～4とパリティ用の送信先ポート5）に変換する。そして、送信先ポート1から4までに接続されているアレモジュール9には分割されたデータが転送され、送信先ポート5に接続されているアレモジュール9には分割されたデータから作成されたパリティが転送される。なお、パリティを作成せず、送信先ポート1から5に接続されている全てのアレモジュール9に分割されたデータを転送することも可能である。

【0058】図7に、DAC4において階層制御ルータ5がRAID3の制御を行う際のフローチャートを示す。図7（a）は正常時の書き込み処理フローチャートを示し、図7（b）は正常時の読み出し処理フローチャートを示し、図7（c）は障害時の読み出し処理フローチャートを示している。

【0059】まず、図7（a）により正常時の書き込み処理を以下に説明する。CPU1より書き込み要求が発生し、シーケンスIDが付けられて、ディスクアレイサブシステム2のDAC4に発行されたとする（ステップ40）。本実施例では、予めDAC4で行うRAID制御のレベル（本例ではRAID3）をユーザが指定しておいてある。そこで、CPU1から発行された、書き込み要求を受け付けた階層制御ルータ5は、予め指定されているRAIDのレベルを認識する（ステップ41）。図7の例ではRAID3の制御を行うように認識する。

【0060】次に、書き込み要求を受け付けた階層制御ルータ5は、キャッシュメモリ7に格納されているルーティングテーブル22を検索する（ステップ42）。本実施例では、RAID3の制御を行うためルーティングテーブル22は図6のようになっている。そして、RAID3の制御をDAC4の階層制御ルータ5が行うため、CPU1から発行された1個の書き込み要求は、図6のルーティングテーブル22により、分割されたデータが書き込まれるアレモジュール9が接続された送信先ポート1～4と、これらの分割されたデータから作成されたパリティが書き込まれるアレモジュール9が接続された送信先ポート5との5個のポートに変換され、

同時にPG6が接続されているポートを認識する（ステップ43）。

【0061】次に、PG6の使用状況を調べ、使用可能な場合は、当該階層制御ルータ5が接続されている上位ポート20とPG6が接続されている上位ポート（PGポート）とを接続する（ステップ44）。当該階層制御ルータ5が接続されているスイッチ8の上位ポート20とPGポートとの接続が完了したら、CPU1から当該階層制御ルータ5およびPGポートを介してPG6へデータを転送し、PG6においてデータの分割を行い、この分割したデータからパリティを生成する（ステップ45）。次に、当該階層制御ルータ5は、PGポートと当該下位ポート21（データ用の送信先ポート1～4とパリティ用の送信先ポート5）を接続するようにPGポートに指示する（ステップ46）。

【0062】このとき、当該アレモジュール9が、他の読み出しまたは書き込み要求で使用中の場合は使用可能になるまで待つ（ステップ47）。当該アレモジュール9が使用可能になったら、これらの当該アレモジュール9が接続されている5個の下位ポートに対し、書き込み要求を発行する（ステップ48）。つまり、階層制御ルータ5は、自分が接続されている上位ポートとPGポートを接続し、PGポート経由でこれらの当該アレモジュール9が接続されている下位ポートに対し、書き込み要求を発行することになる。

【0063】このようにして書き込み要求を受け取った当該アレモジュール9では、アレモジュール9内のMP18が、この要求の受付が可能かどうかを判断し、書き込みが可能な場合は、線19、当該下位ポート21、PGポート、当該上位ポート20を介し階層制御ルータ5に対し転送許可を発行する（ステップ49）。この当該アレモジュール9からの転送許可を階層制御ルータ5が受け取ったら、当該階層制御ルータ5の制御の元で、PG6から当該アレモジュール9へ、分割されたデータおよびこれらのデータから作成されたパリティを転送し、アレモジュール9内では独自に書き込み処理を行う（ステップ50）。

【0064】次に、図7（b）により正常時の読み出し処理を以下に説明する。CPU1より読み出し要求が発生し、シーケンスIDが付けられて、ディスクアレイサブシステム2のDAC4に発行されたとする（ステップ51）。ステップ52、53、54は、上述した正常時の書き込み処理のステップ41、42、43とそれぞれ同じである。

【0065】階層制御ルータ5が複数の送信先下位ポート21とPGポートを認識したら、当該階層制御ルータ5は、PGポートに対し、当該下位ポート21との接続を指示する（ステップ55）。

【0066】当該アレモジュール9が、他の読み出しまたは書き込み要求で使用中の場合は、使用可能になる

まで待つ（ステップ56）。当該アレイモジュール9が使用可能になったら、当該階層制御ルータ5は、自分が接続されている上位ポートとPGポートとを接続し、PGポートを経由してこれらの当該アレイモジュール9が接続されている下位ポートに対し、読み出し要求を発行する（ステップ57）。

【0067】このようにして読み出し要求を受け取った当該アレイモジュール9では、アレイモジュール9内のMP18が、この要求の受付が可能かを判断し、アレイモジュール9内では独自に読み出し処理を行い、CPU1とアレイモジュール9間でデータの読み出しが可能な場合は、線19、当該下位ポート21、PGポート、および当該上位ポート20を介して階層制御ルータ5に対し転送許可を発行する（ステップ58）。この当該アレイモジュール9からの転送許可を受け取った階層制御ルータ5は、当該アレイモジュール9から送出される分割されたデータをPG6へ転送するように制御する（ステップ59）。このようにして各当該アレイモジュール9から受け取った分割されたデータはPG6において結合され、結合されたデータはPG6からPGポートおよび階層制御ルータ5を経由してCPU1へ転送される（ステップ60）。

【0068】次に、図7（c）により障害時の読み出し処理を以下に説明する。まず、本実施例における障害について説明する。本実施例では、アレイモジュール9内のドライブ13の障害はアレイモジュール9内で対策されているものとする。つまり、あるアレイモジュール9内のドライブ13に障害が発生しており、この障害が発生しているドライブに読み出し要求が発行された場合は、アレイモジュール9内の残りの正常なドライブ13に格納されているデータとバリティから、障害ドライブ内のデータを回復して、あたかも正常であるかのようにデータを転送してくる。このように、本実施例のDAC4では、アレイモジュール9内のドライブ障害に対してはアレイモジュール9に任せ、対応しない。

【0069】以下の図7（c）で説明するDAC4の階層制御ルータ5が対応する障害は、アレイモジュール9自身では対応できないような障害である。例えば、アレイモジュール9の制御部に障害が発生し、アレイモジュール9内の全てのドライブ13に対し読み出しまたは書き込みができないような、アレイモジュール9全体に及ぶような障害とする。

【0070】そこで、以下に具体的な処理方法を説明する。

【0071】CPU1より読み出し要求が発生し、シーケンスIDが付けられて、ディスクアレイサブシステム2のDAC4に発行されたとする（ステップ61）。ステップ62、63は、上述した図7（b）の正常時の読み出し処理のステップ52、53とそれぞれ同じである。本実施例では、RAID3の制御を行うためルーテ

ィングテーブル22は図6のようになっている。そして、RAID3の制御をDAC4の階層制御ルータ5が行うため、CPU1から発行された1個の読み出し要求は、図6のルーティングテーブル22により、分割されたデータが書き込まれるアレイモジュール9が接続された送信先ポート1〜4と、これらの分割されたデータから作成されたバリティが書き込まれるアレイモジュール9が接続された送信先ポート5との5個のポートに変換され、同時にPG6が接続されているポートを認識する（ステップ64）。

【0072】なお、アレイモジュール9に障害が発生した場合、ルーティングテーブル22では、障害が発生したアレイモジュール9が接続されている下位ポート名にフラグが付き、階層制御ルータ5はこれによりアレイモジュール9の障害の有無を判断することが可能である。

【0073】次に、PG6の使用状況を調べ、使用可能な場合は、当該アレイモジュール9（障害が発生したアレイモジュール9は除く）が接続されている下位ポート21とPGポートとを接続する（ステップ65）。このとき、当該アレイモジュール9が、他の読み出しまたは書き込み要求で使用中の場合は使用可能になるまで待つ（ステップ66）。当該アレイモジュール9が使用可能になったら、当該階層制御ルータ5は、自分が接続されている上位ポートとPGポートとを接続し、PGポートを経由してこれらの当該アレイモジュール9が接続されている下位ポートに対し、読み出し要求を発行する（ステップ67）。

【0074】このようにして読み出し要求を受け取った当該アレイモジュール9では、アレイモジュール9内のMP18が、この要求の受付が可能かどうかを判断し、アレイモジュール9内では独自に読み出し処理を行い、アレイモジュール9でデータの読み出しが可能な場合は、線19、当該下位ポート21、PGポート、および上位ポート20を介して階層制御ルータ5に対し転送許可を発行する（ステップ68）。この当該アレイモジュール9からの転送許可を受け取った階層制御ルータ5は、当該アレイモジュール9から送出される分割されたデータおよびバリティをPG6に転送するように制御する（ステップ69）。

【0075】PG6では、当該アレイモジュール9から当該下位ポート21およびPGポートを介してPG6へ転送されたデータおよびバリティを用いて、障害が発生したアレイモジュール9内に格納されているデータを回復し、この回復したデータとアレイモジュール9から転送されてきた分割されたデータとを結合する（ステップ70）。このようにして結合されたデータは、PGポートと階層制御ルータ5が接続されている上位ポート20を結合し、階層制御ルータ5を介してCPU1へ転送される（ステップ71）。

【0076】また、もし予備のアレイモジュール9があ

10

20

30

40

50

る場合、または障害が発生したアレイモジュール9を正常なアレイモジュール9に交換した場合は、PG6で回復した障害が発生したアレイモジュール9内のデータを、CPU1ではなく予備のアレイモジュール9または交換した正常なアレイモジュール9に転送し、障害が発生したアレイモジュール9の復元を行うことも可能である。この制御は、所定の階層制御ルータ5が行うようにすればよい。

【0077】(実施例3)

【0078】次に、本発明をRAIDのレベル5(RAID5)に適用した場合を以下に示す。従来技術の欄で説明したが、RAID5では一つのデータを分割せずに1台のドライブに格納する。このとき、ドライブの障害に対する信頼性を向上させるため、各ドライブのデータからパリティを作成し、データとは別のドライブに格納する。このとき、パリティを格納するドライブを特定の1台に固定した場合はRAID4になり、特定せずにデータと同様に複数のドライブに分散させた場合はRAID5になる。RAID5の場合は、データは分割されていないため、ディスクアレイを構成するすべてのドライブが独立に動作することが可能となり、単位時間当りに処理することが可能な読み出し処理件数は向上する。しかし、書き込み時には、パリティを更新するために、書き込まれるアドレスにすでに書き込まれているデータ(旧データ)とパリティ(旧パリティ)を読み出し、新しいパリティを作成した後にデータおよびパリティを書き込む必要がある。したがって、1回の書き込み処理に対して、2回の読み出しと2回の書き込み処理が必要となり、このオーバヘッドが問題となっている。

【0079】RAID5では、先に述べたように書き込み時にはパリティを更新するため、旧データ、旧パリティの2回の読み出しが必要になるため、RAID5の制御を行う場合は、CPU1から書き込み要求が発行されたら階層制御ルータ5は独自に旧データ、旧パリティの2回の読み出し要求を発行する。そこで、以下にその方法を説明する。なお、全体のハードウェア構成、スイッチの内部構成、およびアレイモジュールの内部構成については、上述の実施例1の図1、図2、および図3と同じであるので説明は省略する。

【0080】図8は、RAID5のときのDAC4内の階層制御ルータ5が使用するルーティングテーブル22を示している。各階層制御ルータ5のルーティングテーブル22の使用法は、上述した実施例1、2と同じである。

【0081】RAID5においても、ルーティングテーブル22におけるシーケンスID23に対応するエントリを認識することにより、DAC4内の階層制御ルータ5はRAID制御を行う。本実施例では、DAC4において階層制御ルータ5がスイッチ8を切り替えることでRAID制御を実現する。このため、一つのシーケンス

ID23に対し、階層制御ルータ5がルーティングテーブル22を用いて複数の送信先ポート(図8の例ではRAID5制御を行うためデータの送信先ポートとパリティの送信先ポートの2個のポート)に変換する。

【0082】図9および図10に、DAC4において階層制御ルータ5がRAID5の制御を行う際のフローチャートを示す。図9は正常時の書き込み処理フローチャートを示し、図10(a)は正常時の読み出し処理フローチャートを示し、図10(b)は障害時の読み出し処理フローチャートを示している。

【0083】まず、図9により正常時の書き込み処理を以下に説明する。CPU1より書き込み要求が発生し、シーケンスIDが付けられて、ディスクアレイサブシステム2のDAC4に発行されたとする(ステップ75)。本実施例では、予めDAC4で行うRAID制御のレベル(本例ではRAID5)をユーザが指定している。そこで、CPU1から発行された、書き込み要求を受け付けた階層制御ルータ5は、予め指定されているRAIDのレベルを認識する(ステップ76)。図9の例ではRAID5の制御を行うように認識する。

【0084】次に、書き込み要求を受け付けた階層制御ルータ5は、キャッシュメモリ7に格納されているルーティングテーブル22を検索する(ステップ77)。本実施例ではRAID5の制御を行うためルーティングテーブル22は図8のようになっている。そして、RAID5の制御をDAC4の階層制御ルータ5が行うため、CPU1から発行された1個の書き込み要求は、図8のルーティングテーブル22により、データが書き込まれるアレイモジュール9が接続された送信先ポート173とパリティが書き込まれるアレイモジュール9が接続された送信先ポート274との2個のポートに変換され、同時にPG6が接続されているPGポートを認識する(ステップ78)。

【0085】なお、図8のルーティングテーブル22において、送信先ポート173がデータの送信先ポートであり、送信先ポート274がパリティの送信先ポートである。また、送信先ポート3と送信先ポート4は、送信先ポート274に接続されたアレイモジュール9に格納されているパリティを作成するのに使用したデータが格納されているアレイモジュール9が接続されているポートを示す。つまり、送信先ポート1、3、4に接続されているアレイモジュール9内のデータから作成されたパリティが、送信先ポート2に接続されているアレイモジュール9内に格納されている。したがって、正常時の書き込み処理では、送信先ポート1のアレイモジュール9から旧データを読み出し、送信先ポート2のアレイモジュール9から旧パリティを読み出し、読み出した旧データおよび旧パリティと新規書き込みデータとを用いて新パリティを作成し、新規書き込みデータを送信先ポート1のアレイモジュール9に書き込み、新パリティ

21

を送信先ポート2のアレイモジュール9に書き込むことになる。送信先ポート3、4は後述する障害時に使用する。

【0086】ステップ78の後、階層制御ルータ5は、自分が接続されている上位ポート20とPGポートとをスイッチ8を切り替えて接続する(ステップ79)。次に、CPU1からの書き込みデータを、上位ポート20とPGポートを介してPG6に転送する(ステップ80)。

【0087】RAID5では書き込み時においてパリティを更新するため、旧データ、旧パリティの2回の読み出しが必要になる。このため、RAID5の場合、階層制御ルータ5は、独自に旧データ、旧パリティの読み出し要求を発行する必要がある。以下のステップ81~83は、そのための処理である。

【0088】ステップ78で階層制御ルータ5は既にデータおよびパリティの送信先ポートである下位ポート21を認識しているため、当該階層制御ルータ5は、PGポートと当該下位ポート21(旧データ用の送信先ポート1と旧パリティ用の送信先ポート2)とを接続する(ステップ81)。

【0089】このとき、当該アレイモジュール9が、他の読み出しまたは書き込み要求で使用中の場合は使用可能になるまで待つ(ステップ82)。当該アレイモジュール9が使用可能になったら、旧データおよび旧パリティが格納されているこれらの当該アレイモジュール9が接続されている下位ポートに対し、読み出し要求を発行する(ステップ83)。

【0090】このようにして読み出し要求を受け取った各当該アレイモジュール9では、アレイモジュール9内のMP18が、この要求の受付が可能かを判断し、CPU1とアレイモジュール9間でデータの読み出しが可能な場合は、線19、当該下位ポート21、PGポート、および当該上位ポート20を介して階層制御ルータ5に対し転送許可を発行する(ステップ84)。この当該アレイモジュール9からの転送許可を受け取った階層制御ルータ5は、当該アレイモジュール9から旧データおよび旧パリティを受け取る(ステップ85)。PG6では、このようにして各当該アレイモジュール9から受け取った旧データと旧パリティ、およびステップ80において既にCPU1からPG6に転送されている新データを用いて、新パリティを生成する(ステップ86)。

【0091】次に、当該アレイモジュール9が、他の読み出しまたは書き込み要求で使用中の場合は使用可能になるまで待つ(ステップ88)。当該アレイモジュール9が使用可能になったら、階層制御ルータ5は、これらの当該アレイモジュール9が接続されている下位ポート(送信先ポート1、2)に対し、新データおよび新パリティの書き込み要求を発行する(ステップ89)。

【0092】このようにして書き込み要求を受け取った

22

当該アレイモジュール9では、アレイモジュール9内のMP18が、この要求の受付が可能かを判断し、書き込みが可能な場合は、線19、当該下位ポート21、PGポート、および当該上位ポート20を介して階層制御ルータ5に対し転送許可を発行する(ステップ90)。この当該アレイモジュール9からの転送許可を階層制御ルータ5が受け取ったら、当該階層制御ルータ5の制御の元で、PG6から当該アレイモジュール9へ、新データおよび作成された新パリティを転送し、アレイモジュール9内では独自に書き込み処理を行う(ステップ91)。

【0093】次に、図10(a)により正常時の読み出し処理を以下に説明する。CPU1より読み出し要求が発生し、シーケンスIDが付けられて、ディスクアレイサブシステム2のDAC4に発行されたとする(ステップ92)。ステップ93、94は、上述した正常時の書き込み処理のステップ76、77とそれぞれ同じである。

【0094】正常時の読み出し処理では、CPU1から発行された1個の読み出し要求は、図8のルーティングテーブル22により、データが書き込まれているアレイモジュール9が接続された下位ポート21(送信先ポート1)に変換される(ステップ95)。ステップ95で階層制御ルータ5が当該データが書き込まれているアレイモジュール9が接続された下位ポート21を認識したら、当該階層制御ルータ5が接続されている上位ポート20と当該下位ポート21とを接続する(ステップ96)。

【0095】このとき、当該アレイモジュール9が、他の読み出しまたは書き込み要求で使用中の場合は、使用可能になるまで待つ(ステップ97)。当該アレイモジュール9が使用可能になったら、当該アレイモジュール9が接続されている下位ポートに対し、読み出し要求を発行する(ステップ98)。

【0096】このようにして読み出し要求を受け取った当該アレイモジュール9では、アレイモジュール9内のMP18が、この要求の受付が可能かを判断し、CPU1とアレイモジュール9間でデータの読み出しが可能な場合は、線19、当該下位ポート21、および当該上位ポート20を介して階層制御ルータ5に対し転送許可を発行する(ステップ99)。この当該アレイモジュール9からの転送許可を受け取った階層制御ルータ5は、当該アレイモジュール9から当該データを受け取りCPU1へ転送する(ステップ100)。

【0097】次に、図10(b)により障害時の読み出し処理を以下に説明する。RAID5の障害も、RAID3と同様に、DAC4の階層制御ルータ5が対応する障害はアレイモジュール9自身では対応できないような障害である。そこで、以下に具体的な処理方法を説明する。

50

【0098】CPU1より障害が発生したアレイモジュール9内に書き込まれているデータに読み出し要求が発生し、シーケンスIDが付けられて、ディスクアレイサブシステム2のDAC4に発行されたとする(ステップ101)。ステップ102、103は、上述した正常時の読み出し処理のステップ93、94とそれぞれ同じである。

【0099】RAID5における障害時の読み出し処理では、当該データが作成に関与したバリティと、このバリティ作成に関与した当該データ以外の全データを読み出し、これらから当該データの回復を行う。このため、これらのデータとバリティが書き込まれているアレイモジュール9が接続されている全下位ポート21とPG6が接続されているPGポートを認識する。

【0100】具体的には、ルーティングテーブル22は図8のようになっており、読み出し要求のシーケンスIDに対応する送信先ポート1の下位ポートに接続されているアレイモジュール9に障害が発生している。そこで、当該シーケンスIDに対応する送信先ポート2~4を認識することになる。なお、図8のルーティングテーブル22において、例えば、シーケンスID1で書き込み要求されたデータは下位ポート1に接続されたアレイモジュール9に書き込まれ、シーケンスID2で書き込み要求されたデータは下位ポート3に接続されたアレイモジュール9に書き込まれ、シーケンスID3で書き込み要求されたデータは下位ポート4に接続されたアレイモジュール9に書き込まれ、さらにこれらの書き込みデータから作成されたバリティが下位ポート2に書込まれている。したがって、正常時にはシーケンスID2、3で下位ポート3、4(送信先ポート1)から読み出していたデータを、障害時にはシーケンスID1で下位ポート3、4(送信先ポート3、4)から読み出すことになる。そのため、本実施例では、これらのバリティの作成に用いた複数のデータのシーケンスID(図4)は、異なるシーケンス番号で同じデータアドレスからなるシーケンスIDとし、各アレイモジュール9内ではデータアドレスのみでデータを特定できるようにしてある。したがって、上記シーケンスID1~3の例では、これらのシーケンスID1~3を異なるシーケンス番号で同じデータアドレスから構成されるようにし、正常時にはシーケンスID2、3で下位ポート3、4(送信先ポート1)から読み出していたデータを、障害時にはシーケンスID1で下位ポート3、4(送信先ポート3、4)から読み出すことができるようにしてある。なお、別の方法を用いることもできる。例えば、図8の送信先ポート3、4の欄に当該下位ポートに書き込んだデータのシーケンスIDを書く欄を加えておくようにしてもよい。

【0101】ステップ103の後、CPU1から発行された1個の読み出し要求は、図8のルーティングテーブル22により、バリティの作成に関与したデータとその

バリティが書き込まれているアレイモジュール9が接続されている送信先ポート2~4に変換され、同時にPG6が接続されているポートを認識する(ステップ104)。

【0102】次に、PG6の使用状況を調べ、使用可能な場合は、当該アレイモジュール9が接続されている下位ポート21とPGポートとを接続する(ステップ105)。このとき、当該アレイモジュール9が、他の読み出しまたは書き込み要求で使用中の場合は使用可能になるまで待つ(ステップ106)。当該アレイモジュール9が使用可能になったら、これらの当該アレイモジュール9が接続されている下位ポートに対し、読み出し要求を発行する(ステップ107)。

【0103】このようにして読み出し要求を受け取った当該アレイモジュール9では、アレイモジュール9内のMP18が、この要求の受付が可能かどうかを判断し、アレイモジュール9でデータの読み出しが可能な場合は、線19、当該下位ポート21、PGポート、および上位ポート20を介して階層制御ルータ5に対し転送許可を発行する(ステップ108)。この当該アレイモジュール9からの転送許可を受け取った階層制御ルータ5は、当該アレイモジュール9から送出されるデータおよびバリティをPG6に転送するように制御する(ステップ109)。当該アレイモジュール9より当該下位ポート21およびPGポートを介してPG6へ転送されたデータおよびバリティにより、PG6は、障害が発生したアレイモジュール9内に格納されているデータを回復し、この回復したデータは、PGポートと階層制御ルータ5が接続されている上位ポート20を結合し、階層制御ルータ5を介してCPU1へ転送される(ステップ110)。

【0104】また、もし予備のアレイモジュール9がある場合、または障害が発生したアレイモジュール9を正常なアレイモジュール9に交換した場合は、PG6で回復した障害が発生したアレイモジュール9内のデータを、CPU1ではなく予備のアレイモジュール9または交換した正常なアレイモジュール9に転送し、障害が発生したアレイモジュール9の復元を行うことも可能である。この制御は、所定の階層制御ルータ5が行うようにすればよい。

【0105】(変形例)

【0106】図11は、図1のハードウェア構成の変形例を示す構成図である。図11では図1と異なり、CPU1が階層制御ルータ5を介さずに直接DAC4のスイッチ8の上位ポート20に接続されている。このとき、CPU1からディスクアレイサブシステム2へ読み出しまたは書き込み要求が発行された場合、読み出しまたは書き込み要求を発行したCPU1に接続されている上位ポート20は、必ず階層制御ルータ5が接続されている上位ポート20にスイッチ8を切り換えて接続し、CP

U1からの読み出しまたは書き込み要求は階層制御ルー
タ5で受け付けられる。階層制御ルータ5がCPU1か
らの読み出しまたは書き込み要求を受け付けた後は、図
1と同様に実施例1、2、3で示したようにRAID制
御が行われる。

【0107】以上の実施例では、図1に示すように、D
AC4とn台のアレイモジュール9が1つの筐体に内蔵
されてディスクアレイサブシステム2内を構成している
が、図12に示すように、DAC4の筐体とn台のアレ
イモジュール9が格納されているアレイドライブユニ
ット112の筐体とを分離することも可能である。これ
は、DAC4をバスではなくスイッチ8を中心に構成
し、このスイッチ8に階層制御ルータ5を介してCPU
1やアレイモジュール9を線で接続するようにしたため
である。これにより、DAC4とアレイドライブユニ
ット112とを別の場所に設置することが可能になる。

【0108】また、この拡張として、n台のアレイモジ
ュール9を1ヶ所に集めてアレイドライブユニット11
2の筐体内に収めるのではなく、n台のアレイモジ
ュール9を各々独立に設置することも可能である。このよ
うにすると、n台のアレイモジュール9をそれぞれ別の場
所に設置することが可能になる。

【0109】図13は、複数のDAC4でアレイモジ
ュール9を共有する形態を示す。本発明は、このようにア
レイモジュール9を複数のDAC4が共有した構成にお
いても適用可能である。実現の方式としては、例えば各
アレイモジュールにおけるデータ格納領域をDAC#1
用の領域とDAC#2用の領域との2つに分けておき、
DAC#1からは各アレイモジュールのDAC#1用の
領域にアクセスし、DAC#2からは各アレイモジ
ュールのDAC#2用の領域にアクセスするようにする方式
がある。また、データ格納領域を分けずに、DAC#1
とDAC#2とで同じルーティングテーブルを共有して
(例えば、共通にアクセスできるメモリを設けて、そこ
にルーティングテーブルを共有する)制御するようにし
てもよい。

【0110】このようにアレイモジュール9を複数のD
AC4が共有することにより、DAC4に障害が発生し
た場合や、CPU1とDAC4との間の線113やDA
C4とアレイモジュール9との間の線114に障害が発
生した場合に、もう一方のDAC4やCPU1とDAC
4との間の線113やDAC4とアレイモジュール9と
の間の線114を使用することで、アレイモジュール9
への読み出しまたは書き込みが可能となり、信頼性や可
用性を向上させることが可能となる。また、このように
すると、アレイモジュール9への読み出しまたは書き込
み処理を行えるバスが増加するので、単位時間当りに処
理可能な読み出しまたは書き込み要求数を増加させるこ
とが可能となる。つまり、このアレイモジュール9への
バスがネックになるような場合は、このような構成が有

効となる。

【0111】なお、上記実施例ではキャッシュについて
考慮することなく説明をしたが、図1のキャッシュメモ
リ7に読み出すべきデータがキャッシングされていると
きはキャッシュメモリ7からデータを読み出し、キャッ
シュメモリ7に読み出すべきデータが無い場合に上述の
読み出し処理を行うようにしても良い。また、書き込み
についても、一旦キャッシュメモリ7にデータを書き込
み、後で上述の書き込み処理でキャッシュメモリ7から
ドライブヘデータの書き込みを行うようにしても良い。

【0112】

【発明の効果】複数のディスクアレイモジュール間による
ディスクアレイ制御をクロスバ方式などのスイッチ手
段のルーティング制御により実現することにより、実装
を気にすることなく転送速度が向上する。すなわち、1
本当りの転送速度は従来のバス方式と大差無いが、スイ
ッチを中心にした場合、複数本のバスが並列に動作する
ことが可能となる。したがって、転送速度を向上させる
ことができる。また、スイッチを切り替えることでルー
ティングを行いディスクアレイ制御を行うことで、バス
のようにアービトレーションのような制御が不要になる
ため制御が簡単になる。

【図面の簡単な説明】

【図1】本発明の実施例のハードウェア構成を示す図で
ある。

【図2】図1のDAC内のスイッチの内部構造を示した
図である。

【図3】図1のアレイモジュールの内部構造を示した図
である。

【図4】実施例1のルーティングテーブルを説明する図
である。

【図5】実施例1の読み出しおよび書き込み処理のタイ
ミングチャートを示す図である。

【図6】実施例2のルーティングテーブルを説明する図
である。

【図7】実施例2の読み出しおよび書き込み処理のタイ
ミングチャートを示す図である。

【図8】実施例3のルーティングテーブルを説明する図
である。

【図9】実施例3の正常時の書き込み処理のタイミング
チャートを示す図である。

【図10】実施例3の正常時および障害時の読み出し処
理のタイミングチャートを示す図である。

【図11】実施例の変形例のハードウェア構成を示す図
である。

【図12】実施例の変形例のハードウェア構成を示す図
である。

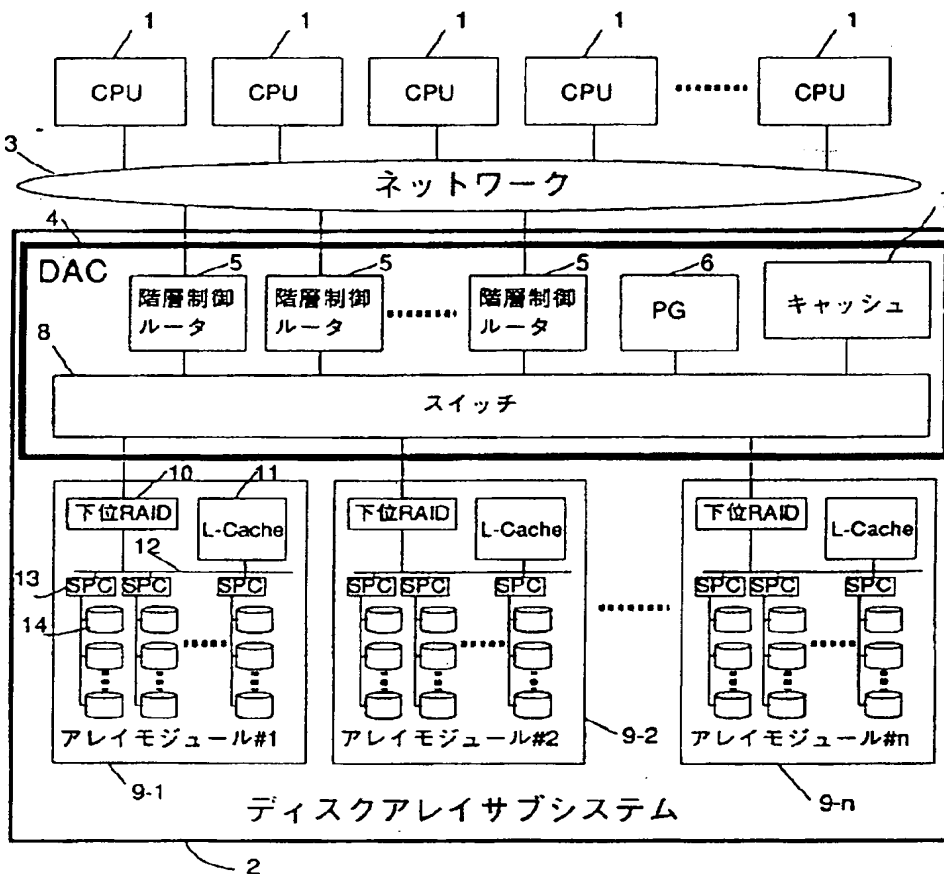
【図13】実施例の変形例のハードウェア構成を示す図
である。

【符号の説明】

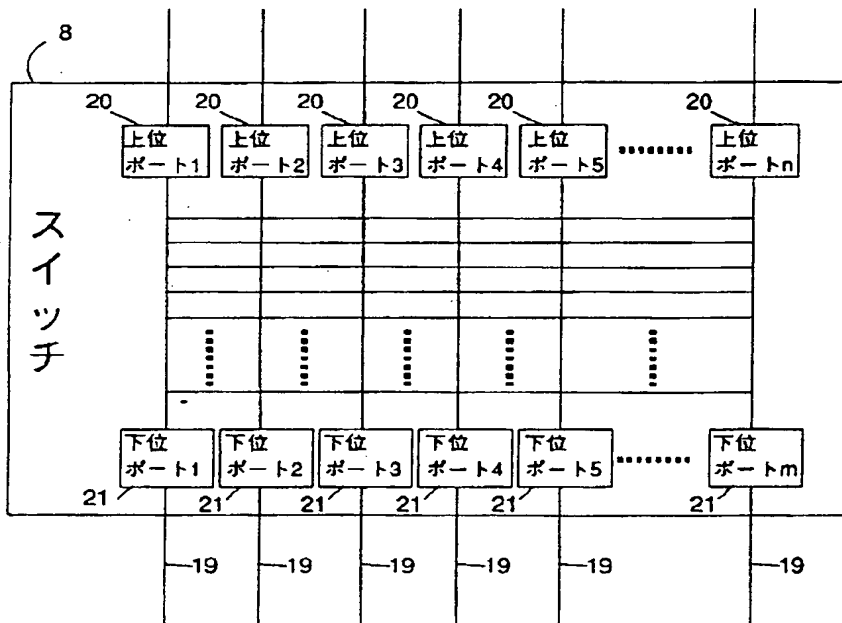
1…CPU、2…ディスクアレイサブシステム、3…ネットワーク、4…ディスクアレイコントローラ(DAC)、5…階層制御ルータ、6…パリティ生成回路(PG)、7…キャッシュメモリ、8…スイッチ、9…アレイモジュール、10…下位RAIDコントローラ、12…内部バス、13…SPC、14…ドライブ、15…MP用メモリ、16…Boot ROM、17…インターフェース制御回路、18…下位RAID制御用MP(M*

*P)、19…線、20…上位ポート、21…下位ポート、22…RAID1用ルーティングテーブル、23…シーケンスID、24…送信先ポート1、25…送信先ポート2、35…RAID3用ルーティングテーブル、72…RAID5用ルーティングテーブル、112…アレイドライブユニット、113…CPUとDAC間の線、114…DACとアレイモジュール間の線。

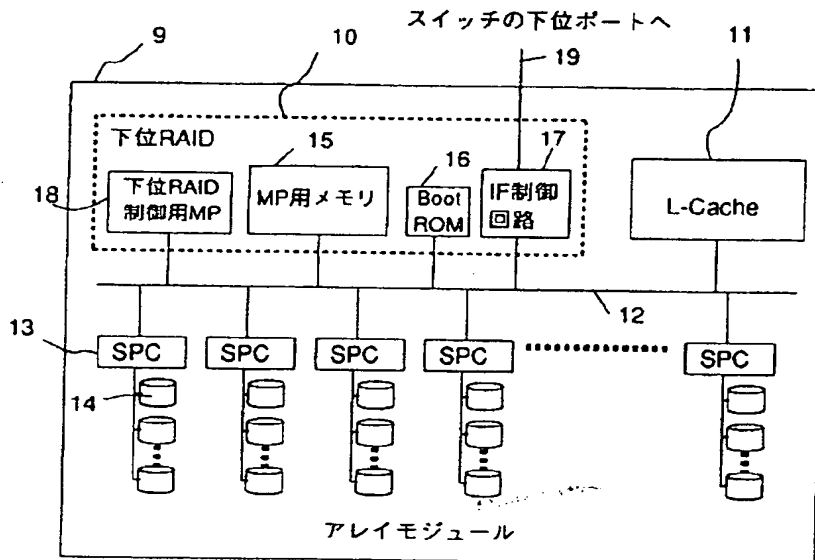
【図1】



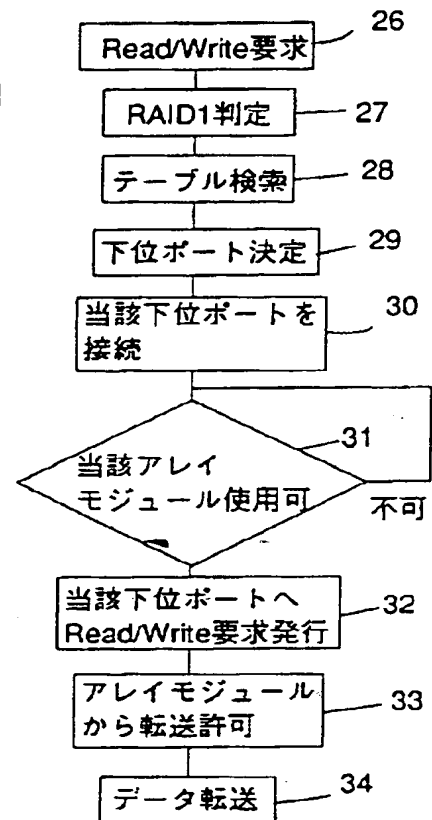
【図2】



【図3】



【図5】



【図4】

シーケンスID		
シーケンス番号	データアドレス	R/Wコマンド

RAID1

シーケンスID	送信先ポート1	送信先ポート2
1	下位ポート1	下位ポート2
2	下位ポート7	下位ポート8
3	下位ポート3	下位ポート4
4	下位ポート5	下位ポート6
⋮	⋮	⋮

【図8】

RAID5(IO Generate)

シーケンスID	送信先ポート1	送信先ポート2	送信先ポート3	送信先ポート4
1	下位ポート1		下位ポート3	下位ポート4
2	下位ポート3	下位ポート2	下位ポート1	下位ポート4
3	下位ポート4		下位ポート1	下位ポート3
4	下位ポート5		下位ポート7	下位ポート8
5	下位ポート7	下位ポート8	下位ポート5	下位ポート8
6	下位ポート8		下位ポート5	下位ポート7
7	下位ポート9		下位ポート11	下位ポート12
8	下位ポート11	下位ポート10	下位ポート9	下位ポート12
9	下位ポート12		下位ポート9	下位ポート11
⋮	⋮	⋮	⋮	⋮

【図6】

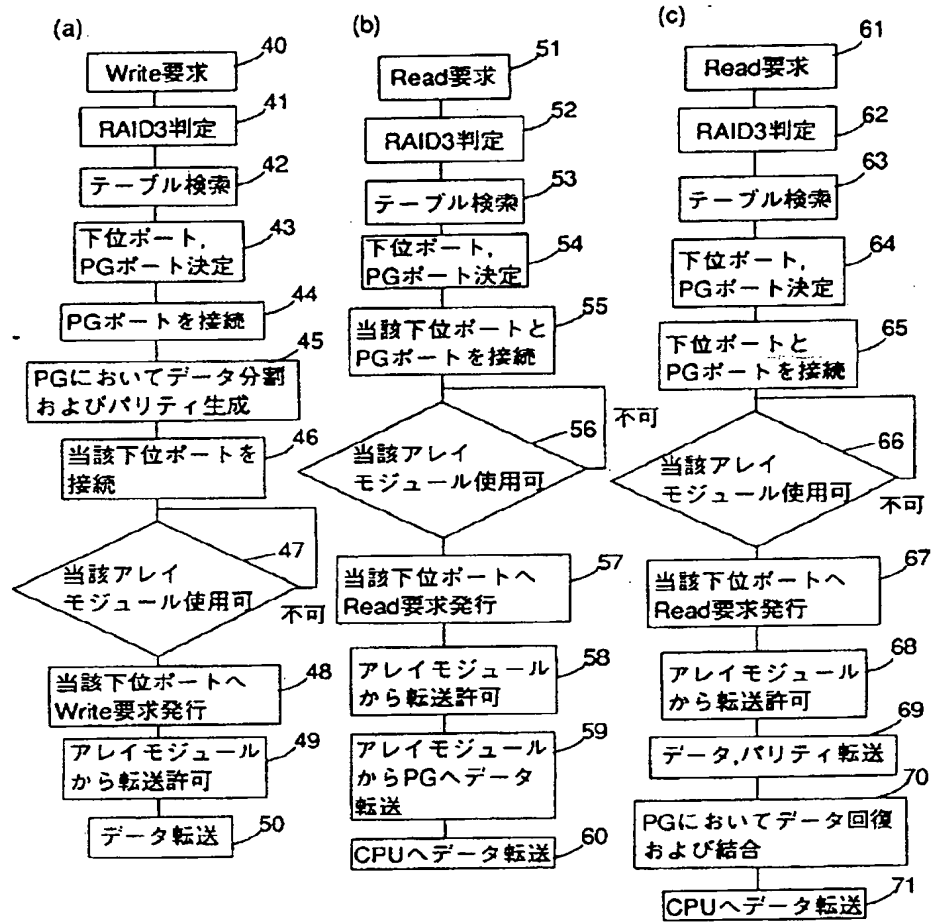
RAID3

シーケンスID	送信先ポート1	送信先ポート2	送信先ポート3	送信先ポート4	送信先ポート5
1	下位ポート1	下位ポート2	下位ポート3	下位ポート4	下位ポート5
2	下位ポート6	下位ポート7	下位ポート8	下位ポート9	下位ポート10
3	下位ポート1	下位ポート2	下位ポート3	下位ポート4	下位ポート5
4	下位ポート6	下位ポート7	下位ポート8	下位ポート9	下位ポート10
⋮	⋮	⋮	⋮	⋮	⋮

【図7】

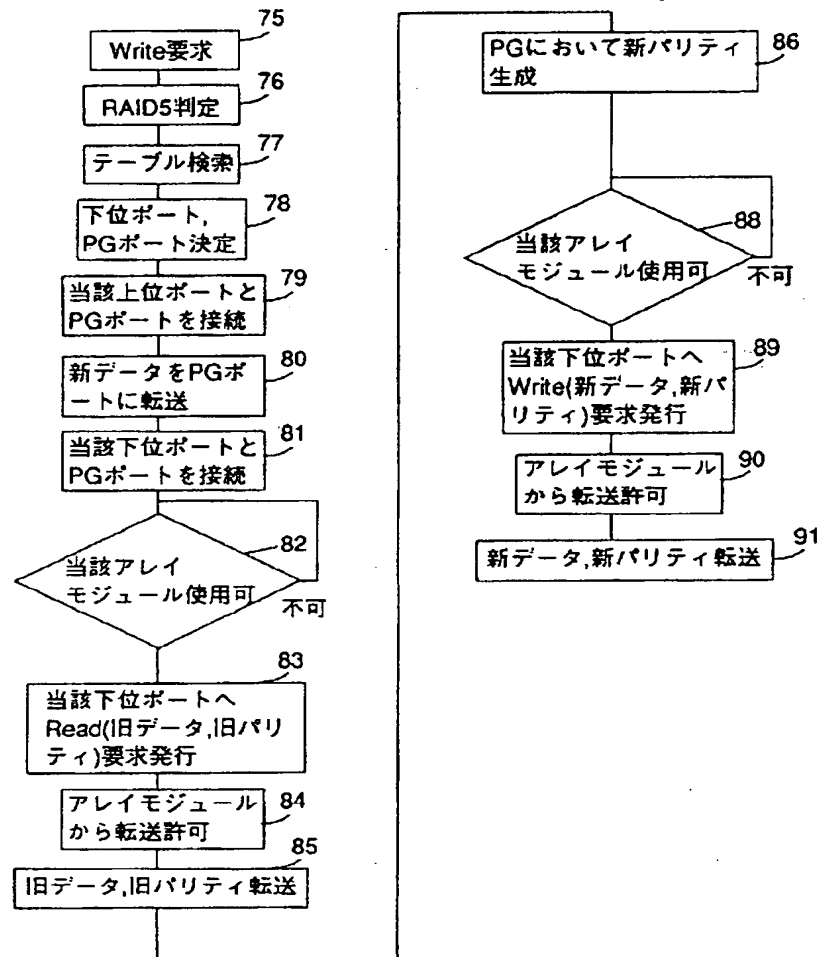
(正常時)

(障害時)

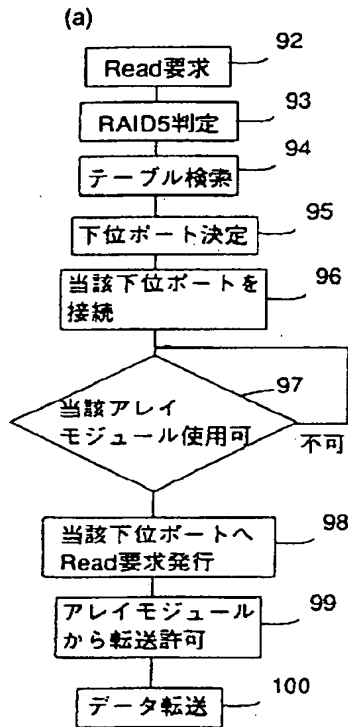


【図9】

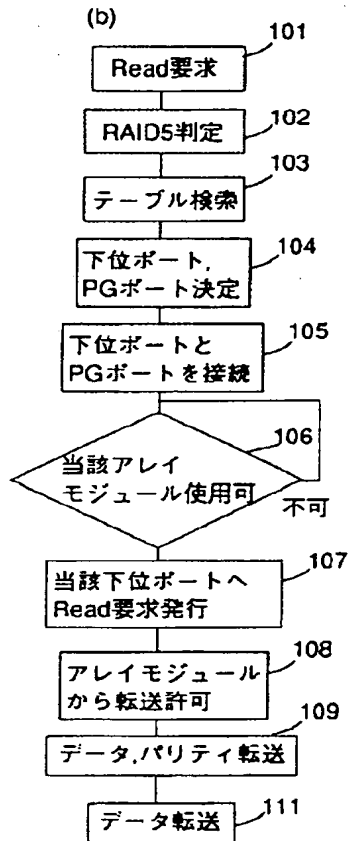
(正常時)



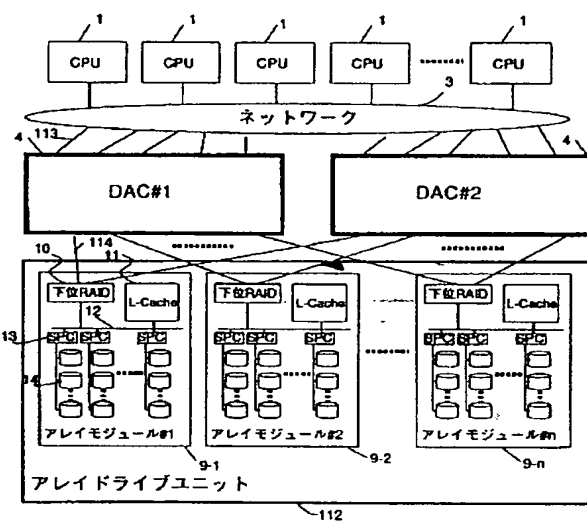
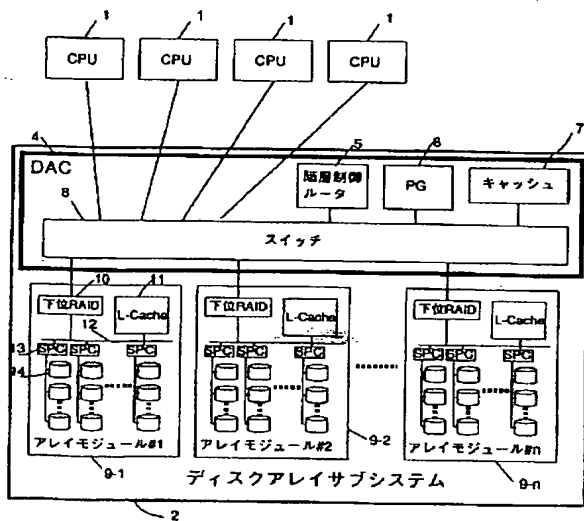
(正常時)



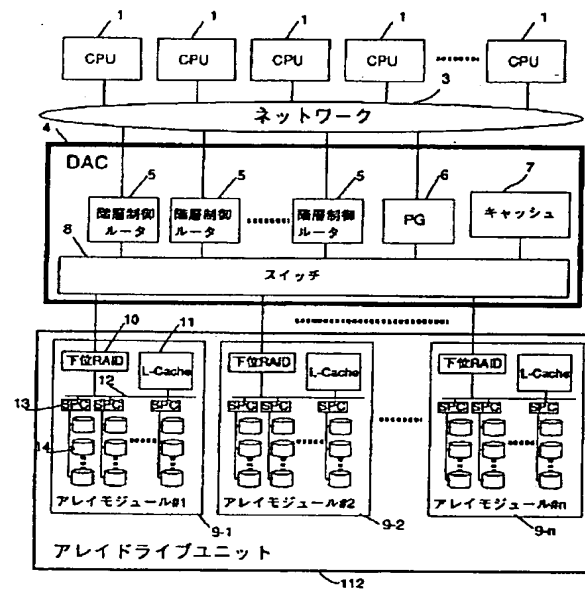
(障害時)



【图 13】



【図12】



フロントページの続き

(72)発明者 加茂 善久
東京都国分寺市東恋ヶ窪一丁目280番地
株式会社日立製作所中央研究所内